humanités numériques

Humanités numériques

3 | 2021
Humanités numériques spatialisées

Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes

Paris in the Travel Writings of Arab Authors: Recognition, Semantic Analysis and Mapping of Toponyms

Motasem Alrahabi, Carmen Brando, Muhamed AlKhalil et Joseph Dichy

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/revuehn/1079
DOI : 10.4000/revuehn.1079
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Motasem Alrahabi, Carmen Brando, Muhamed AlKhalil et Joseph Dichy, « Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes », *Humanités numériques* [En ligne], 3 | 2021, mis en ligne le 01 mai 2021, consulté le 12 mai 2021. URL : http://journals.openedition.org/revuehn/1079 ; DOI : https://doi.org/10.4000/revuehn.1079

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OpenEdition

humanités
numériques

3 | 2021
Humanités numériques spatialisées
INFORMATION SPATIALE ET CORPUS TEXTUELS

Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes

Paris in the Travel Writings of Arab Authors: Recognition, Semantic Analysis and Mapping of Toponyms

Motasem Alrahabi, Carmen Brando, Muhamed AlKhalil et Joseph Dichy

Résumés

À la croisée du traitement du langage naturel, des études littéraires et des humanités spatiales, nous présentons dans cet article une approche pour cartographier les modalités sémantiques positives ou négatives associées aux noms de lieux dans des textes en arabe. La chaîne de traitement comprend le repérage des entités nommées de lieu, l'analyse sémantique de leur contexte (opinions, émotions et sentiments), ainsi que la cartographie de leurs instances sur des cartes géographiques. Notre corpus de travail comprend six récits de voyage à Paris de grands écrivains arabes des XIX^{e} et XX^{e} siècles. Des approches à base de règles et à base d'apprentissage automatique ont été expérimentées et évaluées pour le repérage des entités nommées de lieu et pour l'analyse sémantique. Les résultats de notre étude permettent de confirmer l'apport de cette méthode automatique pour la recherche littéraire, en contribuant à une étude sémantique de vaste ampleur.

We present in this paper an automated method to map out positive or negative semantic modalities associated with place names in Arabic travelogue literature. This research sits at the crossroads of Natural Language Processing, Literary Studies, and Digital Humanities. Our pipeline identifies place named entities, analyzes their semantic context (with regard to opinions, sentiments and emotions), and locates the place

names on geographic maps. Our corpus includes six travel writings on Paris from some of the most influential Arab writers of the 19th and 20th centuries. We evaluate rule-based and machine-learning approaches for their efficacy in named entity recognition and semantic analysis. The results of our automated analysis confirm, to a great extent, the judgements and interpretations of traditional critical scholarship on these Arabic literary texts.

Entrées d'index

MOTS-CLÉS : humanités numériques spatialisées, entités nommées, traitement automatique des langues, analyse sémantique, corpus d'auteur, cartographie, récit de voyage

KEYWORDS: spatial digital humanities, named entities, natural language processing, semantic analysis, authorial corpus, cartography, travel literature

Introduction

1 De nos jours, le nombre de documents en langue arabe dans les bibliothèques et archives nationales du monde entier ne cesse d'augmenter. Cependant, et malgré de récentes initiatives de numérisation [1], ce patrimoine culturel reste souvent peu exploitable sans le recours à des outils intelligents permettant l'analyse sémantique de l'intégralité des textes. L'analyse des entités nommées joue, à juste titre, un rôle très important dans l'interprétation des données littéraires et dans l'accès à leur contenu. Leur traitement automatique consiste généralement à repérer des éléments textuels désignant des objets du monde et à les classer dans des catégories prédéfinies : personnes, lieux, organisations, etc.

2 Il s'agit dans cet article de prendre le repérage d'entités nommées portant sur des lieux comme point de départ, dans l'objectif d'analyser sémantiquement leur contexte : en effet, un lieu ne correspond pas uniquement à une localisation, mais aussi à une construction humaine ; il peut être défini sur le plan des perceptions, des sentiments, des émotions ou des opinions par ceux qui y vivent, qui le traversent ou qui en parlent.

3 À l'articulation du traitement du langage, des études littéraires et des humanités spatiales, notre démarche méthodologique vise à offrir un moyen d'explorer les œuvres littéraires en langue arabe, et plus particulièrement les récits de voyage, en générant de nouvelles connaissances par l'analyse des modalités linguistiques associées aux lieux (opinions, sentiments et émotions).

4 Ce type d'analyse permet de faire des études comparatives des lieux et de leur connotation symbolique selon l'époque, le courant littéraire ou la thématique, de générer des cartographies littéraires (Murrieta-Flores, Donaldson et Gregory 2017) et d'analyser leur évolution en diachronie. Ces analyses permettent aussi de rapprocher des auteurs en fonction des rapports affectifs communs qui les associent à ces différents lieux.

5 Plus précisément, notre étude s'appuie sur l'hypothèse suivante : une analyse des modalités associées aux entités nommées de lieu pourrait constituer un moyen de faire émerger, dans le contexte des études critiques, l'influence de la civilisation française, incarnée par Paris, sur certains écrivains arabes et la façon dont ces écrivains l'ont représentée dans leurs écrits.

6 L'article est organisé en cinq parties. En premier lieu, nous présentons un état de l'art sur l'utilisation de l'analyse spatiale de textes dans les humanités numériques. Dans la deuxième partie, nous expliquons l'importance de Paris dans la culture et la littérature arabes et nous présentons les textes choisis. La troisième partie consiste en une mise en œuvre de notre approche méthodologique : préparation de données, repérage d'entités nommées, analyse sémantique et cartographie. Dans la quatrième partie, nous exposons nos interprétations littéraires des résultats. La dernière partie est consacrée à une discussion au sujet des apports de la chaîne de traitement proposée, mais aussi de ses limites. Nous proposons alors des pistes à explorer afin d'améliorer le travail en cours.

Analyse spatiale de textes et humanités numériques : état de l'art

7 Cette partie a pour objectif de présenter les travaux portant sur la dimension spatiale de textes, plus particulièrement des textes romanesques. En vue de définir et de mettre en œuvre une approche computationnelle, comment définir un lieu ? Comment le désigner, le catégoriser et le localiser dans un espace géographique réel ou abstrait ?

8 Inspiré par la théorie de Bakhtine (1978), le projet *Cartographies chronotopiques* [2] développe une cartographie littéraire de lieux de fictions, en privilégiant la cartographie non référentielle et la cartographie relative. Autrement dit, la carte est un graphe dans lequel les nœuds sont les lieux cités désignant des espaces indéfinis, des lieux de correspondance, des mondes imbriqués ou fantastiques et des espaces d'exil. Les arcs du graphe représentent les distances relatives entre les entités nommées des lieux cités dans le texte. Selon la manière dont ces lieux sont désignés ainsi que le contexte dans lequel ils apparaissent, une série de symboles cartographiques est proposée pour catégoriser les lieux. Par exemple, la catégorie *nature sauvage et idyllique* désigne le monde sauvage, l'ouverture, la liberté, l'espace intact, la terre, le monde naturel, l'unité ; les arcs de type *projection*, par exemple, font référence à un mouvement narratif qui est conduit par l'imagination, la mémoire, les rêves, etc. Une série de cartes sont publiées au fur et à mesure en ligne sur un ouvrage de fiction de la littérature anglaise du XIX^{e} siècle. Par exemple, une carte chronotopique, élaborée à partir du célèbre roman *Frankenstein* [3], montre la dominance du concept d'emprisonnement, car certaines des scènes les plus marquantes concernent des intérieurs clos (le laboratoire et la création de la créature, la chambre à coucher du meurtrier d'Elizabeth…).

9 Aron *et al.* (2017) proposent une analyse chronotopique des *Mystères de Bruxelles* (1845-1846) de Suau de Varennes, afin de faire émerger, à différents niveaux de généralité, les chronotopes structurels d'espèce et de

sous-espèce (roman, roman urbain, « mystères » de Paris et de Bruxelles [4]), et thématiques (*rue huppée*, *commerce à la mode*, *mauvais quartier à assainir*…). À l'aide d'un système d'information géographique, les lieux mentionnés dans cet ouvrage sont cartographiés sur des cartes géographiques contemporaines. Les cartes, précisent les auteurs, proposent une vision synthétique et synoptique de l'espace dans lequel évoluent les personnages. Il s'agit donc de lieux qui peuvent être géolocalisés et projetés sur l'espace géographique réel, certains lieux internes à la fiction (par exemple, le logement d'un personnage) étant également géolocalisés, à condition qu'ils soient situés dans le récit par localisation indirecte (*à proximité de*…). La carte met en évidence l'opposition entre son centre et sa périphérie ainsi que la proximité spatiale des groupes sociaux dans un espace urbain encore faiblement ségrégué.

10 Sur le versant de la nouvelle géographie littéraire, Piatti *et al.* (2009) sont les précurseurs d'une cartographie littéraire de la fiction qui prend appui sur les systèmes d'information géographique (SIG). Les auteurs affirment que les espaces fictifs peuvent comporter, mais sans aucune obligation, des références vers l'espace géographique réel, car lecteurs et auteurs sont tentés par la possibilité d'ancrer d'une manière ou d'une autre les textes dans le monde réel. Les lieux de fiction peuvent être l'endroit précis où l'action se déroule (une maison, un village), une zone regroupant plusieurs décors (une ville entière, une région), les lieux évoqués (en rêve, dans le souvenir et selon une aspiration) ou un itinéraire spatial le long duquel les personnages se déplacent. Ces lieux sont représentés sur la carte avec des emprises floues. Divers romans historiques dans lesquels les actions se déroulent dans les vastes paysages naturels suisses sont analysés dans le but de créer un atlas littéraire européen en ligne [5], influencés par l'*Atlas du roman européen* de Moretti (1999). À partir d'environ 200 textes de fiction, un inventaire détaillé est élaboré avec les lieux cités des trois régions européennes avoisinant la Suisse et l'Allemagne. Plusieurs cartes sur mesure sont réalisées afin de montrer, par exemple, les lieux naturels prédominants chez des auteurs de différentes nationalités.

11 En prenant appui sur les outils de la textométrie et des SIG, plusieurs travaux se sont intéressés à l'analyse historique de l'espace parisien dans les écrits des romanciers français du « grand XIX^e^ siècle ». Moncla, Gaio et Joliveau (2018) et Boeglin (2018) cartographient les empreintes spatiales d'un corpus d'environ 30 romans sur la base de tous les odonymes (noms de voies de circulation) cités dans les textes. Une chaîne de traitement, centrée sur le repérage automatique des odonymes et leur géocodage [6] grâce à un référentiel géohistorique du réseau des voies parisiennes [7], est proposée afin de permettre une exploration interactive et simultanée de l'espace géographique et du texte littéraire. À partir de plusieurs rendus cartographiques, il est possible d'observer les différents quartiers parisiens où s'implantent les histoires des auteurs étudiés (Zola, Balzac) ainsi que la manière dont les empreintes spatiales changent dans le temps, se déplaçant du centre, sur l'île de la Cité, vers la périphérie lorsque la ville de Paris s'élargit. Boeglin (2018) identifie particulièrement quatre incarnations principales de la modernité dans la capitale française : l'architecture, le commerce, les réseaux et les transports, ainsi que les contrastes de la modernité qui écarte les classes les plus démunies de ce nouveau mode de vie.

12 Afin de développer les modèles d'apprentissage machine du traitement automatique des langues (TAL) pour soutenir les études littéraires, Soudani *et al.* (2019) s'intéressent à la reconnaissance d'entités nommées (REN) pour repérer des entités nommées de lieu et de personnes dans la littérature romanesque française du XIX^e^ siècle. Les auteurs de ces travaux insistent sur la difficulté qu'il y a à réutiliser des modèles existants. En effet, ceux-ci ne prévoient pas de textes littéraires dans les corpus d'apprentissage et sont donc peu efficaces pour la REN dans ce type de textes. En vue d'entraîner des modèles, des consignes d'encodage d'entités nommées sont élaborées en s'inspirant de la campagne d'évaluation ESTER 2 [8] et deux modèles REN sont ensuite construits et testés pour deux auteurs (Zola, Balzac) à partir d'un échantillon de romans annotés manuellement. Ces travaux mettent en évidence la difficulté de réutilisation d'un modèle REN entraîné sur un auteur donné. La chaîne proposée s'appuie sur SEM (Dupont 2017) pour la REN, et sur REDEN (Frontini, Brando et Ganascia 2016) pour la désambiguïsation des noms repérés grâce à des référentiels géographiques du Web de données (*Linked Data*). REDEN permet notamment la projection des lieux repérés sur une carte géographique en ligne, car cet outil récupère les informations de géolocalisation du Web de données (DBpedia [9], GeoNames [10]).

13 Dans la même perspective, Borin, Dannélls et Olsson (2014), El Khatib *et al.* (2016) et Does *et al.* (2017) développent des outils et des interfaces d'aide à la cartographie des lieux (réels) cités dans la fiction littéraire par la mise en place d'une chaîne d'outils issus du TAL (adaptés selon la langue du texte) et des SIG. Selon les disponibilités de modèles REN (ici, pour Stanford NER), ces différents outils sont capables de traiter plusieurs langues modernes (anglais, allemand, espagnol, hollandais et suédois).

14 Au delà de la fiction, Frontini *et al.* (2016) proposent une analyse diachronique d'un vaste corpus d'essais littéraires français, publiés entre 1824 et 1932 par des auteurs tels que Bergson, Zola, Sainte-Beuve, Bourget, Faguet, Taine, Brunetière, Lamartine et les Goncourt. L'étude, à une échelle globale, porte sur les nations et les villes citées dans les textes, et tient compte de leur nombre d'occurrences dans le temps. En lisant les différents rendus cartographiques et en maîtrisant le corpus d'étude, ces travaux montrent les lieux de pouvoir et d'influence dans les idées de ces auteurs. Par exemple, la France est la nation la plus citée, ce qui souligne le nationalisme du discours littéraire français ; la Grèce est désignée en tant que berceau de la culture et de la créativité, et l'Italie comme modèle artistique depuis la Renaissance. Du point de vue chronologique, la Russie, peu citée avant 1880, s'impose dans le discours critique entre 1880 et 1900, ce pic correspondant clairement à la découverte de la littérature russe en France.

15 En outre, Gabay et Vitali (2019) s'intéressent au théâtre français du XVII^e^ siècle et proposent une chaîne qui combine les outils du TAL et des SIG, pour analyser des entités nommées de lieu citées dans plusieurs sous-genres (comédie, tragédie et tragicomédie) et à plusieurs périodes historiques. Plusieurs corpus annotés et modèles de TAL issus de l'apprentissage machine sont mobilisés et testés afin de trouver les plus efficaces. Pour le géocodage et la concaténation d'EN, GeoNames est utilisé pour localiser les lieux du début de la modernité et le gazetier Pleiades pour l'Antiquité. Plusieurs analyses littéraires émergent à partir des ren-

dus cartographiques : les auteurs soulignent, par exemple, la dimension mondiale du théâtre français du XVII^e siècle. On y trouve l'Europe, mais aussi les autres continents, avec une concentration de toponymes sur le bassin méditerranéen ; en outre, si on prend l'Europe comme axe de référence, les tragédies sont situées à l'est, les comédies à l'ouest, et les tragicomédies à l'est et à l'ouest.

16 Pour une analyse spatiale de textes tenant compte des émotions associées, le projet de la *cartographie des émotions dans le Londres victorien* [11] illustre la façon dont les romans britanniques des XVIII^e et XIX^e siècles associent les sentiments à diverses parties de la capitale anglaise. Le projet utilise la plateforme *Historypin* et répertorie des passages littéraires, manuellement annotés, qui représentent les quartiers et endroits de Londres sur une carte interactive. Les usagers peuvent ainsi découvrir la façon dont Londres a été imaginée et représentée dans les différentes œuvres de fiction, et quelles émotions lui ont été associées (peur, joie, etc.).

17 Le projet *Mapping the Lake District* [12] (Cooper et Gregory 2011) propose une approche méthodologique entre SIG et linguistique de corpus afin d'explorer les spatialités des récits de voyages de deux poètes anglais, en 1769 et 1802, à travers le paysage du Lake District, dans le nord-ouest de l'Angleterre. Dans ces récits, les paysages naturels du Lake District et de toute l'Angleterre sont qualifiés par ces poètes par des termes tels que *pittoresque*, *magnifique*, *beau*. À l'aide d'un concordancier – outil incontournable de la linguistique de corpus –, il est possible d'analyser les contextes des phrases contenant un nom de lieu et de retrouver les termes appartenant à un lexique d'adjectifs qualificatifs défini au préalable. Ensuite, une série de cartes de chaleur (*heat maps*) sont élaborées par terme grâce à un SIG, afin de visualiser le nombre d'occurrences du terme associées aux différents lieux cités. Ce type de représentation permet d'observer les zones de l'Angleterre qui sont décrites fréquemment par tel ou tel adjectif.

18 Dans un registre textuel différent, Dominguès *et al.* (2019), dans le contexte du projet MATRICIEL, développent une approche intégrant une analyse de lieux et de sentiments sur un corpus de récits de vie de Républicains espagnols exilés en France entre 1936 et 1939 et ayant participé à la Résistance française. Ces travaux mobilisent divers outils du TAL, à la fois des approches à base de règles et à base d'apprentissage automatique, pour le repérage automatique des entités nommées de lieu ; il peut s'agir d'un nom propre pur (*Barcelona*), descriptif (*Jardin des plantes*) ou hybride (*fort du Hâ*), ou d'un nom commun (*zone occupée*, *usine de textile*). Aussi, ces auteurs adaptent un système existant pour l'analyse de sentiments en ajoutant un lexique de termes positifs ou négatifs pertinents dans le contexte des récits de vie. Les règles d'analyse de sentiments tiennent compte de la négation ou des verbes de modalité pour valider ou inverser la polarité du terme trouvé. En matière de cartographie, une carte de densité des entités nommées de lieu (noms propres) pour l'ensemble des récits ainsi qu'une carte soulignant les contrastes des polarités sont élaborées à l'aide d'un SIG.

19 Dans le cadre de notre article, nous proposons une chaîne de traitement qui s'inspire des travaux présentés dans cet état de l'art et qui mobilisent des outils du TAL afin de repérer les entités nommées de lieux

parisiens géolocalisables dans des récits de voyage en langue arabe, et d'attribuer des catégories sémantiques à leur contexte.

Corpus de travail

20 Le récit de voyage est un genre littéraire qui rend compte, selon des modalités diverses, des peuples rencontrés, des émotions ressenties, des choses vues et entendues, ou bien imaginées. Un récit apporte souvent des éléments précieux pour éclairer l'histoire sociale et politique des lieux traversés, voire l'histoire des cultures matérielles, de l'alimentation, des religions, etc. Mais il informe aussi sur la culture du voyageur, sur ses attentes, sur ses préjugés et sur son imaginaire.

21 Tout au long de la Nahda, ou Renaissance arabe, Paris a fasciné les grands écrivains arabes. La capitale de la France est devenue, dans l'imaginaire de millions de lecteurs arabophones, une légende et un symbole de tout ce qui est associé à la civilisation, à la culture, à la prospérité, mais aussi au progrès et à la liberté. Ces derniers termes sont, pour nombre d'auteurs arabes du XIX^e^ siècle et des deux premiers tiers du XX^e^ siècle, auréolés d'une fascination ambivalente, dans laquelle l'attirance pour le progrès et la crainte de ce que ce dernier porte en lui de sulfureux s'entremêlent. De ce fait, nul lieu, à l'époque moderne, ne présente pour la culture arabe et islamique de rencontre aussi forte entre le réel et l'imaginaire que Paris. Cette fascination est liée, au départ, à la rencontre de la modernité, incarnée par cette capitale, ses institutions scientifiques, ses musées, le bouillonnement de ses découvertes mais aussi son mode de vie (Al-Sheikh 1998, Lagrange 2009).

22 Notre choix de textes permet de comparer les différentes représentations de Paris dans l'esprit d'importants écrivains et penseurs arabes à travers une analyse des modalités énonciatives autour des entités nommées de lieu mentionnées dans leurs récits de voyage. Notre sélection couvre des récits de voyage de six grands intellectuels arabes qui ont visité ou vécu à Paris à partir des années 1830 et qui ont investi cette ville comme un ensemble de lieux dont les noms portent en eux une charge symbolique ou poétique qu'il appartiendra à l'analyse littéraire ou historique de dégager. Dans ces œuvres [13], nous étudions comment ces penseurs et écrivains arabes influents, qui ont généralement des convictions idéologiques différentes, ont connu Paris à différentes époques, et comment ils ont dépeint la ville dans leurs écrits, entre admiration et critique [14].

23 Le premier ouvrage que nous avons choisi est le livre fondateur تخليص الإبريز في تلخيص باريز (*L'Or de Paris*) de Rifa'a al-Tahtawi (1801-1873), un imam et écrivain égyptien. Ce récit de voyage relate les observations et commentaires d'al-Tahtawi lors de son séjour à Paris entre 1826 et 1831 lorsqu'il accompagna, en tant qu'imam, de jeunes Égyptiens envoyés se former à différentes sciences à Paris par Mohammad Ali Pasha, souverain d'Égypte. Ses écrits et sa vision de la vie civique et culturelle française ont donné le ton à une nouvelle attitude libérale envers l'Occident dans le monde arabe.

24 Le deuxième ouvrage est le livre كشف المخبّا عن فنون أوروبا (*La Découverte des voies cachées des arts d'Europe*) écrit par Ahmad Faris Al-Shidyaq (1804-1887). Dans ce récit de voyage, Al-Shidyaq couvre plusieurs villes et

pays, mais nous avons choisi son deuxième chapitre, qu'il consacre à Paris. Al-Shidyaq était un éditeur, poète, essayiste, lexicographe et traducteur pionnier ainsi que l'un des pères fondateurs du journalisme arabe. C'était un réformiste et un progressiste inébranlable, traversé par des diversités religieuses. Il était également connu pour être anglophile, ce qui ajoute une dimension supplémentaire à ses vues sur Paris et la France.

25 Le troisième ouvrage est le livre الدنيا في باريس (*L'Univers à Paris*) d'Ahmad Zaki Pasha (1867-1934), récit de voyage relatant son séjour en France en 1900. Zaki Pasha était un important philologue et homme politique égyptien. D'où l'importance du livre, écrit par un fonctionnaire du gouvernement, un traducteur parlant couramment le français et un nationaliste, devenu l'un des principaux intellectuels arabes du tournant du XX^{e} siècle.

26 Le quatrième ouvrage est le récit de voyage رحلة إلى أوروبا (*Un voyage en Europe*) de Jurji Zaydan (1861-1914), dans lequel nous n'avons choisi que la section couvrant la France et Paris (environ la moitié du livre). Zaydan est un écrivain et journaliste libanais qui a établi, presque à lui seul, une fiction historique présentant une vision héroïque de la civilisation islamique arabe et jetant les bases de ce qui deviendra la conscience panarabe. En 1882, Zaydan a fondé *al-Hilal*, revue littéraire et culturelle mensuelle influente, toujours publiée aujourd'hui.

TABLEAU 1. CORPUS DE TRAVAIL

Auteur	Titre	Date	Nombre de mots
رفاعة رافع الطهطاوي Rifa'a al-Tahtawi	تخليص الإبريز في تلخيص باريز *L'Or de Paris*	1834	70 231
أحمد فارس الشدياق A. Fares Al-Shidyaq	كشف المُخبَّا عن فنون أوروبا *La Découverte des voies cachées des arts d'Europe*	1857	25 870 (sur 105 743)
أحمد زكي Ahmad Zaki	الدنيا في باريس *L'Univers à Paris*	1900	50 271
جرجي زيدان Jurji Zaydan	رحلة إلى أوروبا *Voyage en Europe*	1912	16 163 (sur 26 052)
زكي مبارك Zaki Mubarak	ذكريات باريس *Souvenirs de Paris*	1931	52 690
مالك بن نبي Malek Bin Nabi	مذكرات شاهد للقرن *Mémoires d'un témoin du siècle*	1965	50 389

27 Le cinquième ouvrage est le livre ذكريات باريس (*Souvenirs de Paris*) de Zaki Mubarak (1892-1952), un mémoire sur la vie et les observations de l'auteur pendant son année d'étude à Paris à partir de juillet 1930. Mubarak était un savant azharite qui s'aventura au delà cette éducation traditionnelle en suivant une haute éducation française. Après avoir obtenu son doctorat en littérature à l'Université égyptienne en 1924, il a reçu un diplôme de l'École nationale des langues orientales vivantes en 1931 puis un doctorat de l'université de la Sorbonne en 1937. Mubarak était un écrivain prolifique : il est l'auteur de 45 livres, dont deux en français.

28 Le dernier ouvrage est le livre مذكرات شاهد للقرن (*Mémoires d'un témoin du siècle*) de Malik Bin Nabi (1905-1973), qui était un intellectuel islamique algérien. Bien que ce livre ait été publié en 1965, la moitié de celui-ci traite directement des difficultés de vie et d'éducation de l'auteur à Paris

entre 1930 et 1954. Bin Nabi, réformiste islamique et conservateur, offre une vision différente de Paris, souvent montrée de manière plus critique dans le contexte de la lutte algérienne pour l'indépendance et des clivages civilisationnels plus larges entre le monde musulman et le monde occidental.

29 Le tableau 1 regroupe les différentes œuvres de notre sélection de textes pour l'analyse.

Quelques particularités de la langue arabe

30 Différentes difficultés peuvent être observées au niveau du traitement automatique de la langue arabe, notamment l'absence totale ou partielle des signes diacritiques (Habash 2010). Le rôle de ceux-ci est initialement de lever l'ambiguïté pour la lecture et l'interprétation ; cependant, leur utilisation n'est pas systématique dans la plupart des textes aujourd'hui. L'absence de lettres capitales dans le système orthographique de la langue arabe rend difficile la reconnaissance des noms propres, contrairement à d'autres langues où la capitalisation est une caractéristique importante pour les systèmes de REN. Ajoutons à cela le fait qu'une bonne partie des noms propres arabes ont la forme d'adjectifs nominalisés (كريم, أنور, مصطفى, هيفاء) et deviennent autonymiques en perdant leur valeur prédicative (Dichy 2004).

31 Enfin, le phénomène d'agglutination dans une langue flexionnelle comme l'arabe fait que certaines particules, parfois même combinées, peuvent s'attacher aux noms propres, ce qui multiplie les cas d'ambiguïté autour d'une EN, notamment en l'absence quasi-systématique des signes de voyellation dans les textes arabes. Citons à titre indicatif les prépositions et les articles de coordination (proclitiques), les possessifs ou les marques du pluriel (enclitiques).

Reconnaissance des entités nommées de lieu

32 Notre démarche consiste à repérer les noms de lieux puis à annoter les modalités linguistiques dans leur contexte, avant de projeter les résultats sur des cartes.

33 En TAL, les noms de lieux sont généralement étudiés dans le cadre des entités nommées. Une entité nommée (EN) est une expression linguistique dénotant un référent unique d'un domaine quelconque (Nouvel, Ehrmann et Rosset 2015). Elle peut évoquer des personnes, des lieux, des organisations, des expressions temporelles, ainsi que toute autre catégorie du monde réel ou imaginaire.

34 Selon Dominguès *et al.* (2019), géographes et linguistes s'accordent à dire que « la matière est généralement divisée selon la géographie » en montagnes, rivières et cours d'eau, lieux habités, etc. Dans ce sens, un lieu correspond à une portion de territoire située sur la terre, et peut être désigné par son nom, souvent répertorié dans les *gazetiers géographiques*[15] (IGN[16], GeoNames, Pleiades[17]). De par leurs définitions, leurs propriétés et leurs usages, les noms propres sont généralement bien adaptés pour désigner des lieux (Jonasson 1994).

35 La reconnaissance d'entités nommées est une étape essentielle pour de nombreuses tâches en TAL, comme la fouille de textes, la traduction automatique ou la recherche d'information. En témoignent les nombreuses campagnes d'évaluation (MUC-6 et MUC-7, CoNLL, ACE, ESTER-2 [18]) et les abondants travaux sur différentes langues (Nadeau et Sekine 2007, Sharnagat 2014, Shaalan 2014, Etaiwi, Awajan et Suleiman 2017).

36 Outils existants pour la REN en arabe

37 Comme pour les autres langues, les systèmes de REN de l'arabe ont été développés en utilisant principalement trois approches : approches linguistiques à base de règles manuellement créées (Mesfar 2008, Shaalan et Raza 2009, Ben Mesmia *et al.* 2017) ; approches statistiques ou à base d'apprentissage automatique avec des données pré-étiquetées (Benajiba *et al.* 2007, Pasha *et al.* 2014, Abdelali *et al.* 2016, Helwe et Elbassuoni 2017) ; approches hybrides combinant les deux approches précédentes (Oudah et Shaalan 2012, Alotaibi et Lee 2014).

38 Parmi les outils disponibles récents, nous avons commencé par tester et évaluer les trois systèmes de REN suivants :

- Stanza (Qi *et al.* 2020) est une librairie en Python qui utilise le moteur CoreNLP de l'université Stanford. La boîte à outils est conçue pour plusieurs dizaines de langues, dont l'arabe, en utilisant le formalisme des dépendances universelles.
- Farasa (Abdelali *et al.* 2016) est une suite d'outils pour le traitement de textes en arabe qui propose, entre autres, un module de REN basé sur un apprentissage semi-supervisé et des ressources multilingues de *Wikipédia* (Darwish 2013).
- Madamira (Pasha *et al.* 2014) est également une suite d'outils pour le traitement de langue arabe : analyse morphosyntaxique, diacritisation, annotation des parties de discours, REN, etc.

39 Afin de mener notre évaluation, nous avons procédé à la constitution d'un corpus de référence (*gold standard*).

Corpus de référence

40 Nous avons manuellement annoté les entités nommées de lieu du premier tiers de chacun des six livres. À l'aide de la plateforme INCEpTION (Klie *et al.* 2018 [19]), l'annotation a été effectuée par trois étudiants arabophones de niveau universitaire. Les annotateurs ont suivi un protocole d'annotation que nous avons élaboré à cette fin, en nous inspirant du modèle ESTER-2 [20]. Ainsi, pour refléter la granularité utilisée par nos auteurs dans la description des espaces, nous avons défini quatre types d'entités nommées de lieu :

- Lieux naturels et géographiques (Loc-Nature) : نهر السين/*la Seine* ; جبال الألب/*les Alpes*…
- Régions administratives (Loc-Admin) : الحي اللاتيني/*Quartier latin* ; باريس/*Paris*…
- Bâtiments et constructions fonctionnelles (Loc-Building) : المتحف الحربي الفرنسي/*musée français de l'armée* ; دار الأوبرا/*l'Opéra*…
- Chemins et axes de trafic (Loc-Path) : ساحة الكونكورد/*place de la Concorde* ; بولفار سان ميشيل/*boulevard Saint-Michel*…

41 Le *gold standard* a été vérifié par calcul d'accord entre annotateurs avec une concordance importante de 0,78 selon la mesure de Kappa Fleiss. Les divergences constatées concernaient souvent des désaccords entre les deux sous-catégories Admin et Nature (*Île-de-France*), ou entre Building et Path (*Porte de Vincennes*). Les lieux inconnus des annotateurs pourraient être également une source de désaccords (ديوان البير/*la Chambre des pairs ;* مارستان المجانين/*l'hôpital des fous ;* كرسي مملكة الفرنسيس/*capitale du royaume de France*). Le tableau suivant répartit les 532 EN uniques retenues après le calcul d'accord [21] :

TABLEAU 2. LES EN RETENUES DANS LE *GOLD STANDARD*

Auteur	Loc-Admin (≈ 23 en moyenne)	Loc-Building (≈ 55 en moyenne)	Loc-Nature (≈ 3 en moyenne)	Loc-Path (≈ 8 en moyenne)
Rifa'a al-Tahtawi	28	108	3	0
Malek Bin Nabi	30	55	5	24
Zaki Mubarak	20	35	3	4
Ahmad Zaki	20	99	3	21
Jurji Zaydan	22	32	0	0
A. Fares Al-Shidyaq	17	0	3	0

42 Le guide d'annotation ainsi que le *gold standard* sont librement accessibles en ligne [22].

43 Pour la comparaison des trois systèmes, nous avons choisi le livre de Rifa'a al-Tahtawi (1834). Cet ouvrage, le plus ancien de notre corpus, a la particularité de nommer, souvent pour la première fois en arabe, un grand nombre de lieux en France et à Paris, par des procédés de traduction littérale, d'arabisation et de translittération (Aly Mohamed Aly 2012). Le tiers annoté de ce livre contient 149 phrases, 8 741 *tokens* et 362 annotations manuelles.

Évaluation des trois systèmes d'apprentissage de REN

44 Les trois systèmes permettent d'identifier les catégories couramment utilisées – Personne, Organisation et Lieu –, mais seule la dernière nous intéresse. Aucun des systèmes ne propose une granularité fine pour la catégorisation des lieux. Nous avons donc gardé de notre corpus de référence uniquement le premier niveau d'annotation (Loc) en éliminant les spécifications Admin, Nature, Building et Path.

45 Une autre difficulté a été observée lors de la comparaison : les trois outils proposent une annotation au format BIO, courant dans les travaux de TAL, mais ne fournissent pas un découpage comparable au niveau du mot (la phase de *tokenisation*, ou segmentation). Afin d'éviter toute source d'erreur, nous avons vérifié et ajusté manuellement l'alignement des trois sorties d'annotation.

46 Une procédure d'évaluation automatique [23] a été mise en œuvre pour calculer les performances en termes de précision (P), de rappel (R) et de F-mesure (moyenne harmonique entre le rappel et la précision) [24]. Les résultats, calculés aux niveaux des correspondances exactes et des correspondances partielles, sont présentés dans le tableau 3 :

TABLEAU 3. LES RÉSULTATS D'ÉVALUATION DES TROIS SYSTÈMES

	Mesure	Partielle	Exacte
Farasa	Précision	0,54	0,42
	Rappel	0,45	0,35
	F1	0,49	0,38
Stanza	Précision	0,51	0,41
	Rappel	0,72	0,58
	F1	0,59	0,48
Madamira	Précision	0,35	0,25
	Rappel	0,55	0,39
	F1	0,43	0,31

47 Nous remarquons que les meilleures performances en termes de F-mesure sont affichées par le système Stanza de Stanford, avec un rappel relativement important par rapport aux deux autres systèmes. Madamira occupe la deuxième position en termes de rappel alors que Farasa a le meilleur score en précision parmi les trois systèmes.

48 Signalons que ces systèmes REN sont entraînés ou évalués sur un type particulier de textes, généralement modernes, comme les articles de journaux ou *Wikipédia* (Alanazi 2017), et que l'évaluation de leur performance pourrait se heurter à la variabilité des données, de sorte que les résultats sur d'autres genres de corpus restent souvent décevants (Soudani *et al.* 2019). Ce manque de robustesse à la variation est particulièrement criant dès lors qu'il est question des corpus littéraires (variations diachroniques, diatopiques…).

49 Dans ce sens, nous envisageons d'entraîner un nouveau modèle NER à granularité fine avec Stanza. Cependant, vu la taille actuelle du corpus annoté [25], il semble plus concret d'explorer une solution provisoire qui fournit des annotations précises pour la suite de la chaîne de traitement (analyse des modalités et cartographie).

50 Ainsi, à partir de notre corpus de référence, nous avons constitué un gazetier et mis en place un système à base de règles pour annoter l'ensemble du corpus.

Annotation fine des EN à l'aide d'un gazetier

51 À partir des EN collectées dans le corpus de référence, nous avons créé un gazetier pour les noms de lieux appartenant à la France, sans aucun enrichissement lexical. Les différentes formes agglutinées (proclitiques et enclitiques) de chaque entrée ont été systématiquement prises en compte. Le tableau suivant récapitule la distribution, par catégorie et par auteur, des 2 111 EN identifiées avec cette approche (valeurs non uniques) :

TABLEAU 4. LES EN RECONNUES AVEC UNE APPROCHE À BASE DE DICTIONNAIRE

Auteur	Loc-Admin	Loc-Building	Loc-Nature	Loc-Path
Rifa'a al-Tahtawi	390	159	39	0
Malek Bin Nabi	307	40	4	5
Zaki Mubarak	418	89	31	4
Ahmad Zaki	107	30	9	15
Jurji Zaydan	205	17	0	0
A. Fares Al-Shidyaq	223	17	2	0

52 Nous remarquons que la catégorie Loc-Admin prédomine avec environ 80 % de l'ensemble des occurrences reconnues, devant la catégorie Loc-Building. Pourtant, celle-ci contient plus d'EN dans le *gold standard* (voir tableau 2).

FIGURE 1. RÉPARTITION DES EN RECONNUES SUR L'ENSEMBLE DU CORPUS

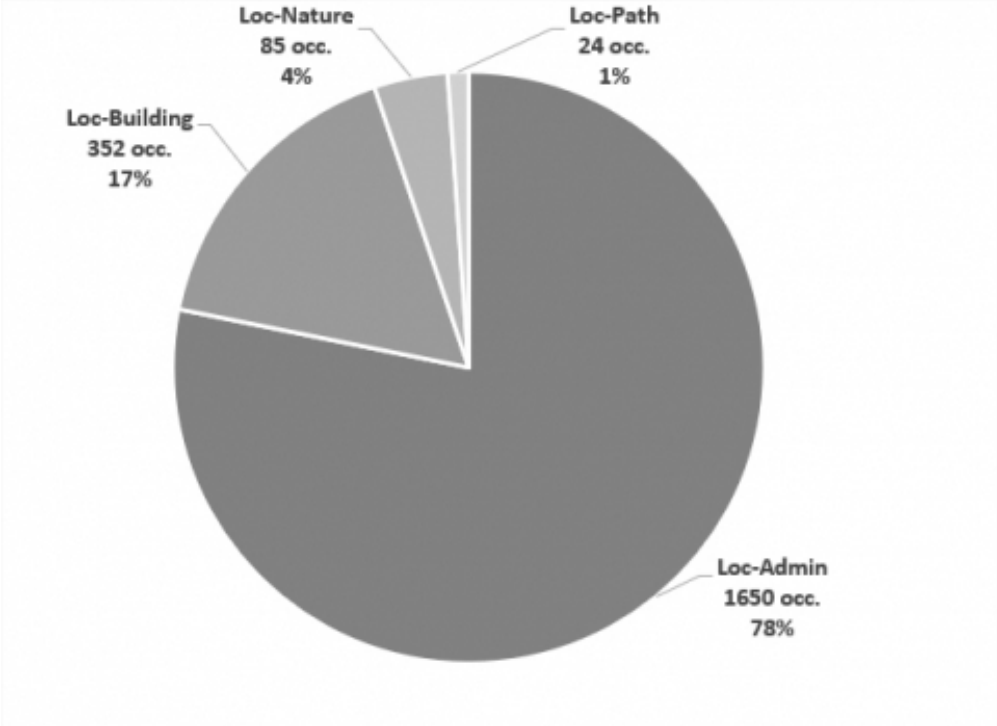

Image produite par les auteurs

53 Voici les dix EN les plus fréquentes pour chaque livre :

TABLEAU 5. LES EN RECONNUES LES PLUS FRÉQUENTES, PAR AUTEUR (APPROCHE À BASE DE DICTIONNAIRE)

Auteur	Les 10 EN reconnues les plus fréquentes
Rifa'a al-Tahtawi	باريس, فرنسا, السين, ديوان رسل العمالات, ديوان البير, مرسيليا, بيت المال, ليون, ديوان الملك, بستان النباتات *Paris*, *France*, *la Seine*, *Chambre des députés*, *Chambre des pairs*, *Marseille*, *Trésor public*, *Lyon*, *la Cour du Roi*, *Jardin des plantes*
Malek Bin Nabi	باريس, الحي اللاتيني, فرنسا, مرسيليا, فرساي, معهد الدراسات الشرقية, باب سان دونيس, باب فرساي, شارع تريفيز, ليون *Paris*, *Quartier latin*, *France*, *Marseille*, *Versailles*, *Institut d'études orientales*, *Porte de Saint-Denis*, *Porte de Versailles*, *rue de Trévise*, *Lyon*
Zaki Mubarak	باريس, فرنسا, الحي اللاتيني, مرسيليا, السين, السوربون, حديقة النباتات, لكسمبور, قهوة الجامع, حديقة لكسمبور

	Paris, *France*, *Quartier latin*, *Marseille*, *la Seine*, *la Sorbonne*, *Jardin des plantes*, *Luxembourg*, *Café de la mosquée*, *Jardin du Luxembourg*
Ahmad Zaki	فرنسا, نهر السين, مارسيليا, مدينة پاريس, التروكاديرو, ميدان شان دومارس, برج إيفل, باريس, معرض پاريس, قصر التروكاديرو *France*, *la Seine*, *Marseille*, *Paris*, *Trocadéro*, *Champ-de-Mars*, *tour Eiffel*, *Exposition de Paris*, *palais du Trocadéro*
Jurji Zaydan	باريس, فرنسا, ليون, بوردو, مرسيليا, طولوز, بواتيه, قصر اللوفر, كليني, نوتردام *Paris*, *France*, *Lyon*, *Bordeaux*, *Marseille*, *Toulouse*, *Poitiers*, *Palais du Louvre*, *Cluny*, *Notre-Dame*
A. Fares Al-Shidyaq	باريس, فرنسا, التولري, كالي, مارستان السقط, الأويرة, صان جرمان, قصر الملك, نهر السين, بستان النباتات *Paris*, *France*, *les Tuileries*, *Calais*, *hôpital des Invalides*, *l'Opéra*, *Saint-Germain*, *Palais royal*, *la Seine*, *Jardin des plantes*

54 Les entités nommées les plus citées dans l'ensemble des œuvres sont, sans surprise, *Paris* et *France*, suivis par *Marseille*, *la Seine*, *Lyon*, *Jardin des plantes* et *Quartier latin*.

Évaluation de l'approche par dictionnaire

55 Après les premiers tests, nous avons observé que la difficulté majeure rencontrée avec cette méthode symbolique relevait de l'ambiguïté contextuelle (Leidner 2008). À titre d'exemple, le mot ليل renvoie à la ville Lille ou bien au mot *nuit* ; le mot السين signifie *la Seine* ou bien le nom de la lettre « s » en arabe ; le mot كان peut être la ville de Cannes ou bien l'auxiliaire *être* au passé, très fréquent. Afin de surmonter cette difficulté, nous avons ajouté aux entrées ambiguës des indices contextuels comme les termes génériques introduisant les EN de lieu : *rue*, *avenue*, *place*, etc. La deuxième difficulté concerne ce qui est appelé la surcomposition, c'est-à-dire la présence d'une EN dans deux catégories différentes (Moncla et Gaio 2015). Par exemple, dans un passage qui contient حديقة لوكسمبورغ/*le Jardin du Luxembourg*, le système annote une fois *Luxembourg* en tant que Loc-Admin (*la ville*) et une deuxième fois *le Jardin du Luxembourg* en tant que Loc-Building. La solution proposée est d'ignorer les EN faisant partie d'une séquence déjà annotée.

56 Afin d'évaluer l'approche par dictionnaire, nous avons choisi de tester les motifs sur un nouveau récit de voyage, *Est et Ouest* (شرق وغرب), écrit par Muhammad Husayn Haykal (1888-1956). Nous avons sélectionné les chapitres qui relatent ses voyages à Paris, ce qui représente 182 phrases (environ 6 000 mots). Les résultats obtenus contiennent 154 EN dans 81 phrases annotées, avec une précision de 100 % et un rappel de 81 %. Les 36 EN non reconnues sont des lieux non répertoriés ou bien des lieux appartenant à notre dictionnaire mais sous une transcription différente : نهر الصين au lieu de نهر السين et فنسين au lieu de فانسين...

57 Rappelons que le gazetier provient du premier tiers de chacun des six livres du corpus. Un enrichissement de cette base donnerait nécessairement de meilleurs résultats.

Analyse des modalités linguistiques autour des EN de lieu

58 Après l'étape de REN, nous procédons à l'analyse de leur contexte afin d'annoter les modalités linguistiques, avec leur polarité positive ou négative. Cela nous permet de mettre en évidence la répartition spatiale des parcours de vie des auteurs, déterminés par le contexte historique de chaque expérience.

59 En TAL et en fouille de données, les travaux relatifs aux modalités recoupent souvent l'analyse des opinions, des sentiments, des émotions et des points de vue (Turney 2002, Wiebe *et al.* 2004, Pang et Lee 2008, Zhang et Liu 2017, Schiller, Daxenberger et Gurevych 2020). La tâche classique de ce domaine consiste à déterminer la polarité globale d'une opinion, d'un sentiment ou d'un point de vue, émis par une source et portant sur une cible. Il est de tradition dans ces travaux de distinguer les faits des opinions. Dans le premier cas, il s'agit de descriptions objectives ; dans le second, il s'agit d'expressions subjectives ou évaluatives, qui peuvent porter par ailleurs sur des entités descriptibles objectivement (Boullier et Lohard 2012).

60 Certains travaux dans ce domaine se basent sur l'utilisation de lexiques ou dictionnaires regroupant des termes avec leurs valences : SentiWordNet (Esuli et Sebastiani 2006), EMOTAIX (Piolat et Bannour 2009). Mais les modalités peuvent être véhiculées par d'autres éléments linguistiques au niveau de la morphosyntaxe ou du discours, et les phénomènes de la négation (Zhang 2012) et de l'intensité y jouent un rôle très important.

61 En partant de certains travaux linguistiques comme ceux de Charles Bally (1932), Antoine Culioli (1999) ou Jean-Pierre Desclés et Zlatka Guentcheva (2000), nous distinguons au sein de toute phrase deux parties : la représentation (*dictum* ou relation prédicative) et la modalité (*modus* ou participation active que le sujet opère sur le *dictum*). Dans l'exemple *Heureusement, Pierre est guéri*, le marqueur *heureusement* indique une modalité appréciative que l'énonciateur porte sur le reste de l'énoncé.

62 Contrairement au repérage d'EN, nous avons opté dès le départ pour une annotation des modalités par une approche fondée sur le lexique. En effet, nous avons réutilisé des ressources linguistiques existantes pour le repérage des modalités énonciatives dans les textes arabes (Alrahabi 2010, Alrahabi 2015). Ces ressources linguistiques, principalement verbales, ont été enrichies par des adjectifs et des adverbes et classées en motifs de surface, dans une trentaine de catégories :

- catégories positives : accord, appréciation, joie, soutien… (603 marqueurs au total, regroupés dans 13 catégories)
- catégories négatives : accusation, colère, tristesse, désaccord… (777 marqueurs au total, regroupés dans 24 catégories)

63 Certaines structures de modalités qui génèrent du bruit ont été écartées, et l'intensité n'a pas été traitée dans le présent travail. Cependant, nous avons pris en compte dans la construction des ressources linguis-

tiques les particules de la négation syntaxique (لا, لم, لن) qui peuvent complètement inverser la polarité modale.

64 À l'aide d'un outil à base de règles, Excom-2 (Alrahabi 2010), nous avons procédé à la reconnaissance automatique de ces motifs dans l'ensemble du corpus. Une phase de prétraitement automatique est nécessaire pour chaque texte : structuration au format XML-TEI, ajout de métadonnées (auteur, date, éditeur…) et, enfin, segmentation en phrases délimitées par des signes de ponctuation (point, point-virgule, points d'exclamation ou d'interrogation). Cette dernière étape est nécessaire pour délimiter les espaces de recherche dans lesquels s'effectue le repérage des marqueurs linguistiques. Ce repérage s'effectue avec des règles contextuelles faisant appel à des marqueurs de surface, sans analyse morphologique préalable [26].

65 Seules les phrases contenant à la fois des EN et des annotations sémantiques ont été conservées, à savoir 692 phrases. L'outil y a identifié 993 EN et 1 311 modalités. Parmi ces dernières, 785 annotations sont « positives » (60 %) et 526 sont « négatives » (40 %). Le tableau suivant récapitule l'ensemble des résultats :

TABLEAU 6. LES CATÉGORIES SÉMANTIQUES IDENTIFIÉES DANS LE CONTEXTE DES EN

Polarité	Étiquette	Occ.	Polarité	Étiquette	Occ.
Négative	Accusation	12	Positive	Accord	45
Négative	Ambiguïté	2	Positive	Appréciation	660
Négative	Avertissement	19	Positive	Assurance	3
Négative	Colère	5	Positive	Correct	5
Négative	Critique	13	Positive	Guérison	15
Négative	Dénonciation	2	Positive	Joie	1
Négative	Dépréciation	422	Positive	Paisible	2
Négative	Désaccord	9	Positive	Pardonner	2
Négative	Incorrect	4	Positive	Résultat_Positif	20
Négative	Indignation	5	Positive	Excuses	4
Négative	Insulte	4	Positive	Soins	17
Négative	Ironie	7	Positive	Soutien	11
Négative	Mépris	1			
Négative	Plainte	4			
Négative	Résultat_Négatif	8			
Négative	Tristesse	8			
Négative	Vantardise	1			

66 Voici quelques exemples de sorties :

إنها لرياح ذلك الزمن، الرياح التي كانت تصرف إلى باريس كل جزائري تخفق أحلامه وتفشل مشاريعه في بلادي.

C'est le vent de cette époque, le vent qui envoyait à Paris tous les Algériens dont les rêves et les projets échouaient dans mon pays. (Bin Nabi)/Négatif

ومما يبهر العقول في باريس دكاكين الكتبية وخاناتهم، وتجارات الكتب، فإنها من التجارات الرائجة مع كثرتها وكثرة المطابع، وكثرة التآليف التي تنطبع كل سنة فإنها يعسر حصرها

Ce qui éblouit à Paris, ce sont les librairies et leurs boutiques, le commerce des livres, qui est très répandu, avec les nombreuses librairies et imprimeries, et la création littéraire qui est imprimée chaque année, si abondante qu'on ne peut la compter. (al-Tahtawi)/Positif

فأما ما في باريس من الصروح الفاخرة والمباني السنية فمما لا يعد ولا يحصى، ولكني أذكر منها أشهرها، فمن ذلك القصر المسمى باللوفر.

Les édifices somptueux et les bâtiments illustres à Paris sont innombrables, mais je ne mentionne que les plus célèbres parmi eux, par exemple le palais appelé le Louvre. (Al-Shidyaq)/Positif

فلما كانت سنة ١٨٧٨ أقامت فرنسا معرضا عامًا كبيرا، وبقي منه إلى الآن قصر التروكاديرو الجميل

En 1878, la France a organisé une grande exposition universelle, et il en reste à ce jour le magnifique palais du Trocadéro. (Zaki)/Positif

فالعلة الأصلية في شيوع التهتك بباريس، إنما هو إطلاق سراح الفتاة ومساواتها للرجل، وتكليفها الارتزاق مثله، وإباحة الحكومة للفحشاء رسميًا، وزد على ذلك أن الفتور الديني شائع في فرنسا

La raison première de la prévalence du libertinage à Paris est la liberté accordée à la femme et son égalité avec l'homme, qu'on lui demande de gagner sa vie comme lui, et la légalisation officielle de l'obscénité par le gouvernement. De plus, l'apathie religieuse est courante en France. (Zaydan)/Négatif

أكتب إليك هذه الرسالة من روان مدينة الماضي والأحلام والفن الجميل

Je vous écris cette lettre de Rouen, la ville du passé, des rêves et du bel art. (Mubarak)/Positif

67 Nous trouvons par la suite la répartition des catégories positives et négatives chez chaque auteur :

FIGURE 2. RÉPARTITION PONDÉRÉE DES ANNOTATIONS SÉMANTIQUES POSITIVES ET NÉGATIVES CHEZ CHAQUE AUTEUR

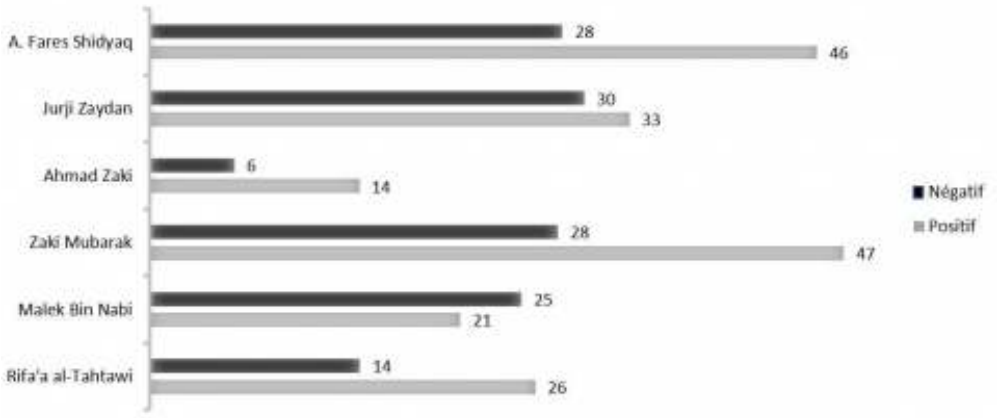

Image produite par les auteurs

68 Dans ce graphique, le nombre d'annotations par auteur a été pondéré par le nombre total de mots de l'œuvre en question, puis multiplié par 10 000 pour la clarté des proportions.

Évaluation du repérage automatique des modalités

69 Le processus d'annotation sémantique a été également évalué. Le texte de M. Haykal a été traité manuellement par l'un de nos annotateurs afin d'attribuer à chacune des 81 phrases contenant des EN une ou plusieurs des valeurs suivantes : positive, négative, les deux à la fois, aucune des deux. L'annotateur a laissé 59 annotations. Nous avons ensuite comparé ces annotations manuelles avec celles obtenues automatiquement par Excom-2 sur le même texte : il y a 36 bonnes annotations, 10 mauvaises annotations et 13 annotations manquées[27]. Si la précision (78 %) est bonne, le rappel (61 %) reste relativement faible, à cause notamment de marqueurs non encore répertoriés, comme la phrase suivante qui a été annotée par l'outil comme Positive, alors qu'il s'agissait également d'une modalité négative dont le marqueur ne faisait pas partie de nos ressources : فقد كدت أضيق ذرعا بباريس رغم حبي إياها/*J'en avais presque assez de Paris malgré mon amour pour elle*. Notons également que notre approche se base uniquement sur les marqueurs observables. Certains phénomènes, qui ne sont pas repérés par les règles, peuvent nous échapper, comme l'ironie, l'allusion ou le sous-entendu.

Cartographie

70 La cartographie vise la représentation des lieux évoqués dans les récits de voyage ainsi que les modalités qui y sont associées. Elle permet de mettre en évidence la répartition spatiale des différents récits à l'aide de représentations bimodales associant texte et carte et permettant de proposer des interprétations pertinentes. Le processus de cartographie que nous proposons consiste à retrouver les coordonnées géographiques des entités nommées de lieu identifiées et à les projeter sur des cartes thématiques avec les modalités linguistiques associées. Comme nous l'avons déjà souligné, à ce stade nous nous intéressons uniquement aux lieux réels et géographiquement attestés. Le travail de géocodage a été effectué manuellement, afin d'attribuer à chaque nom de lieu reconnu par le système des coordonnées géographiques sur une carte. Plusieurs cas de figures et des difficultés ont été rencontrés :

- Les lieux qui n'existent plus : ديوان البير/*la Chambre des pairs*… Nous avons décidé d'associer leur localisation historique à leur lieu actuel et d'afficher le nom tel qu'il apparaît dans le livre.
- Les lieux dont l'orthographe ou la transcription a changé : الشمزليزه/*Champs-Élysées* ; أكدمة/*Académie* ; لوندرة/*Londres* ; باريز/*Paris*… Nous avons regroupé les variantes d'un lieu autour de la même entrée.
- Les lieux dont les noms ont complètement changé au cours de l'histoire : مارستان/*maristan*, aujourd'hui hôpital ; كرسي/*siège*, aujourd'hui *capitale*… Nous avons associé à ces noms la nouvelle localisation, tout en affichant sur la carte les noms mentionnés dans le livre.
- Les EN ambiguës, qui renvoient à des lieux en France ou ailleurs : قبر الجندي المجهول/*tombe du soldat inconnu* ; أكدمية الفلسفة/*l'Académie de philosophie*… Nous avons gardé ces lieux en leur attribuant une localisa-

tion en France, après vérification.

- Les lieux collectifs non déterminés : موانئ فرنسا/*les ports de France* ; مارستانات باريس/*hôpitaux de Paris* ; المستعمرات الفرنسية/*les colonies françaises*… Pour ce travail, nous avons écarté ces lieux de la carte.

71 Les figures 3 et 4 sont les cartes de Paris de deux auteurs éminents de notre corpus, al-Tahtawi et Bin Nabi, présentant les polarités associées aux entités nommées de lieu repérées dans leurs livres. Un premier regard sur ces cartes permet de constater que la perception de al-Tahtawi de Paris est très positive. En revanche, il n'y a pas de tendance forte vers une polarité ou une autre pour Bin Nabi (voir la partie « Interprétation des résultats »).

FIGURE 3. LES LIEUX PARISIENS CITÉS PAR AL-TAHTAWI ET LEURS POLARITÉS (LE PLUS GRAND ANNEAU CORRESPOND À PARIS)

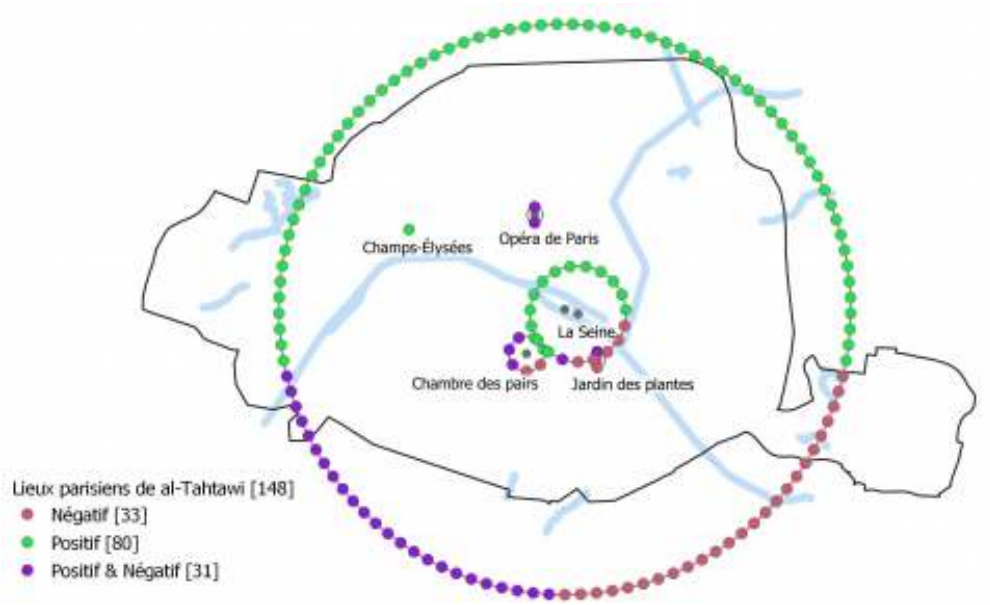

Image produite par les auteurs

72 La « Seine » reste un lieu particulièrement apprécié par al-Tahtawi. Sur la rive droite, les lieux de al-Tahtawi sont historiques, sans connotation clairement négative. Bin Nabi cite une grande diversité de lieux, qui s'étendent du centre de la ville au Quartier latin, jusqu'à ses portes (*portes de Vincennes*, *de Saint-Denis* et *de Versailles*). En effet, les moyens de transport (routier, ferré) sont déjà répandus à l'époque, il est donc possible de se déplacer sans contraintes par les axes parisiens et le métro. Certaines voies sont appréciées par Bin Nabi comme *la rue Saint-Jacques* et *la rue de Trévise*, et d'autres très dépréciées comme *la rue Leconte*. Pour les deux auteurs, on trouve aussi quelques lieux importants pour la culture arabe et orientale, *la mosquée de Paris* et *l'Inalco*.

FIGURE 4. LES LIEUX PARISIENS CITÉS PAR BIN NABI ET LEURS POLARITÉS (LE PLUS GRAND ANNEAU CORRESPOND À PARIS)

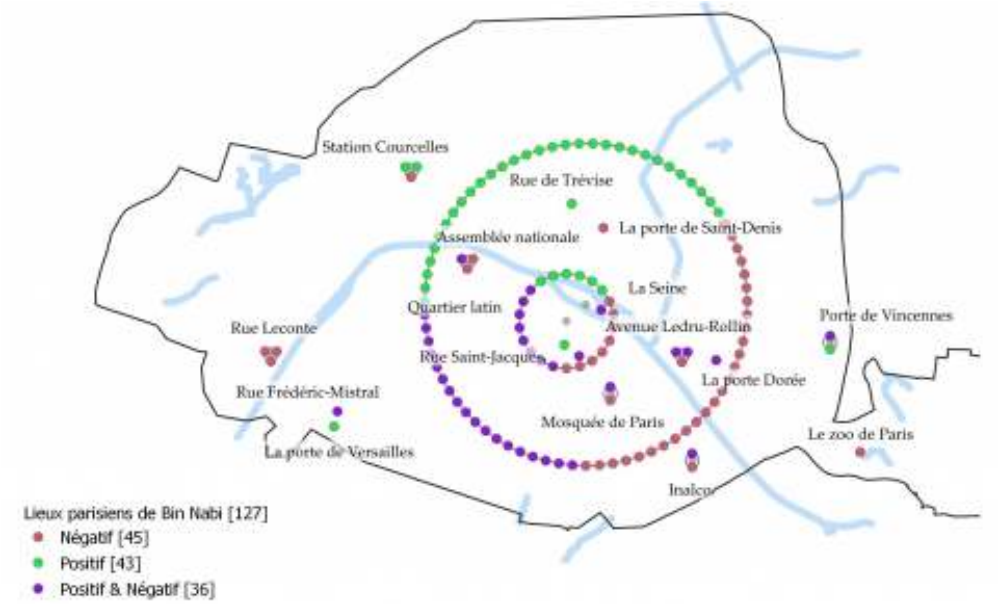

Image produite par les auteurs

73 Nous avons également réalisé une carte interactive, qui peut être consultée en ligne [28], où nous avons projeté les entités nommées de lieu de tout le corpus sur une carte géographique à l'aide de l'application *Google Maps*.

FIGURE 5. UNE CARTOGRAPHIE EN LIGNE DE TOUS LES LIEUX IDENTIFIÉS À PARIS À POLARITÉ POSITIVE, NÉGATIVE, MIXTE OU NEUTRE

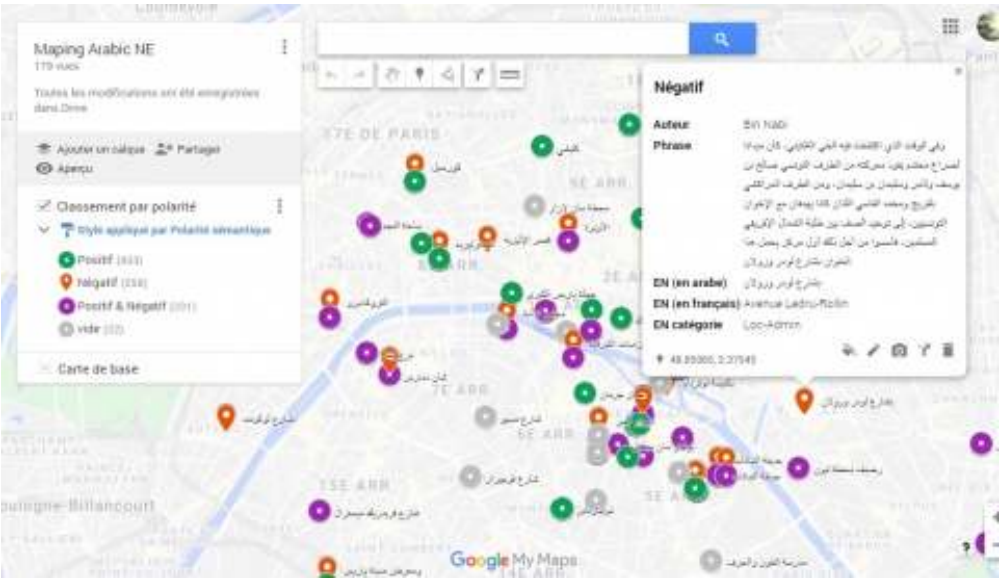

Image produite par les auteurs

74 La densité spatiale des lieux, cartographiés selon leur catégorie (administrative, bâtiments, axes ou nature), donne une lecture synthétique de la géographie des récits dans notre corpus. La polarité, représentée en couleurs, apporte de nouvelles interprétations quant à la manière selon laquelle les auteurs se sont approprié les lieux visités ou mentionnés.

Interprétation des résultats

75 Dans la sélection des œuvres à analyser, nous avons voulu comparer les résultats obtenus par nos outils avec des études critiques disponibles sur le thème de notre recherche. Les critiques arabes ont récemment commencé à analyser et à évaluer les écrits clés sur Paris dans la littérature arabe moderne, abordant cette ville comme le nœud central dans la relation dialectique entre les civilisations arabe et française modernes. Ces études récentes concernent les écrits des intellectuels et hommes de lettres arabes qui ont visité Paris au XIX^e^ et au début du XX^e^ siècle, et examinent comment ils ont rendu compte de leurs expériences dans la littérature qu'ils ont produite. Parmi les critiques importants qui ont travaillé sur la question, citons Khalil Al-Sheikh, qui a publié en 1998 son livre *Paris in Modern Arabic Literature*. Il y examine les écrits arabes en prose sur Paris jusqu'au milieu du XX^e^ siècle, ainsi que des écrits poétiques composés après cette date [29].

76 Dans son introduction, Ihsan Abbas note que l'ensemble des intellectuels arabes qui sont entrés en contact pour la première fois avec la civilisation occidentale à travers Paris étaient en conflit, divisés entre deux points de vue opposés : celui de la ville de la science, de la connaissance et du développement et celui de la ville profane, ville des péchés [30]. Khalil Al-Sheikh note la même division mais moins en tant que conflit que comme une dualité dans l'expérience de Paris, entre, d'une part, l'effervescence et le plaisir de la découverte de la civilisation française, d'autre part, le découragement et la prise de conscience de la civilisation arabe et de son « retard » perçu.

77 Parcourant les écrits arabes sur Paris, Al-Sheikh identifie en outre deux thèmes-clés, qu'il qualifie de « dimensions ». Le premier est le thème du *paradis*, dérivé des imaginations religieuses, qui, par exemple dans les écrits d'al-Tahtawi et d'Ahmad Zaki, imprègne la ville d'idéalisme et de sainteté. Le deuxième thème qu'il note est celui de la civilisation, Paris devenant le symbole d'une renaissance. Ce motif civilisationnel apparaît plus clairement dans les écrits d'auteurs comme Taha Hussein (1889-1973), qui ont visité ou étudié à Paris après la Première Guerre mondiale et qui ont préconisé l'adoption du modèle français pour parvenir au renouveau arabe. Dans notre sélection d'écrivains, Zaki Mubarak fait également partie de ce groupe qui regarde très positivement Paris et le mode de vie français. Dans les résultats de notre analyse automatique, les trois auteurs – al-Tahtawi, Zaki et Mubarak – ont présenté des pourcentages de sentiments et d'émotions similaires, où les modalités positives étaient presque deux fois plus nombreuses que les modalités négatives. On peut également ajouter à ce groupe Ahmad Faris Al-Shidyaq, qui, comme le note Al-Sheikh, « n'est pas venu à Paris d'une ville arabe mais de Londres, une ville qui rivalise avec elle en développement. C'est pourquoi Al-Shidyaq n'a pas sanctifié et vénéré Paris autant que d'autres qui venaient de l'Est arabe. » Cependant, l'observation d'Al-Sheikh n'a pas été confirmée dans notre analyse automatique, car Al-Shidyaq présente des pourcentages de sentiments similaires aux trois visiteurs en provenance des pays arabes. Il est possible qu'Al-Sheikh

ait tiré sa conclusion de ce qu'Al-Shidyaq a écrit dans la totalité de son livre, y compris de son expérience à Londres, finalement décevante, et que nous avons exclue de notre corpus.

78 En opposition à ce projet revivaliste arabe et francophile viennent des projets revivalistes antifrançais construits sur des bases panarabes ou panislamiques. Al-Sheikh note l'expérience entièrement différente de l'intellectuel et visionnaire islamique Malik Bin Nabi, qui a vu Paris comme une sorte d'enfer. Bin Nabi est d'abord arrivé à Paris en tant que travailleur algérien, ce qui l'a exposé à un niveau de difficultés, d'exploitation et de préjugés que n'ont généralement pas connu les autres écrivains privilégiés de l'Est arabe. Cela a coloré sa vision de la ville et de la civilisation française en général, ce qui l'a conduit à concevoir un projet anticolonial antagoniste de renouveau panislamique. Sa vision très négative de Paris et de la France se confirme en effet dans notre analyse automatique de ses écrits. Bien qu'il ait encore des sentiments positifs envers Paris, il est le seul auteur de notre sélection pour lequel notre système a rapporté des modalités plus négatives que positives.

79 Les écrits de Jurji Zaydan sur Paris n'étaient pas inclus dans le livre d'Al-Sheikh, malgré le rôle critique que Zaydan a joué dans la formation de la conscience historique des Arabes modernes. À bien des égards, Zaydan est l'ancêtre littéraire du nationalisme arabe et il était enrichissant de l'inclure dans notre étude pour voir comment des écrivains comme lui, avec leur fierté d'appartenir à la civilisation arabe, regardaient Paris. Bien qu'il n'ait pas présenté une expérience aussi négative que Bin Nabi, Zaydan est toujours le seul écrivain de notre groupe à avoir des sentiments presque aussi négatifs que positifs, ce qui peut suggérer que la montée du nationalisme arabe a contribué à une vision plus prudente de la civilisation française.

80 Indépendamment des termes utilisés pour décrire les tropes que les auteurs arabes ont utilisés pour décrire Paris, il ressort clairement de l'étude d'Al-Sheikh que si les opinions des auteurs arabes sont divisées sur Paris, dans l'ensemble, elles sont beaucoup plus positives que négatives. Cela correspond parfaitement à nos résultats d'analyse : dans l'ensemble, notre système a détecté dans ces écrits 60 % de modalités positives contre 40 % de modalités négatives. Ces chiffres confirment à la fois la division conflictuelle et l'impact plus positif de l'expérience arabe à Paris.

Interface graphique pour l'exploration des résultats

81 Afin d'illustrer les résultats de notre approche, nous avons intégré les textes annotés dans une interface générique, intitulée Ariane [31], développée dans le cadre du Labex OBVIL. Cette application permet de naviguer dans les segments préalablement annotés selon notre ontologie de modalités (Alrahabi 2015). En l'occurrence, dans le cadre de ce projet, il est possible de retrouver tous les passages portant des étiquettes de modalités (voir le tableau 6), d'y rechercher des mots-clés ou éventuellement des EN. L'utilisateur a également le moyen de croiser les résultats avec les métadonnées « auteur », « date », etc. À titre d'exemple, il est possible de formuler la requête suivante : quels sont les passages portant une connotation négative et contenant le nom de lieu *Paris* dans l'œuvre de Ben Nabi ?

FIGURE 6. ARCHITECTURE GÉNÉRALE DU SYSTÈME

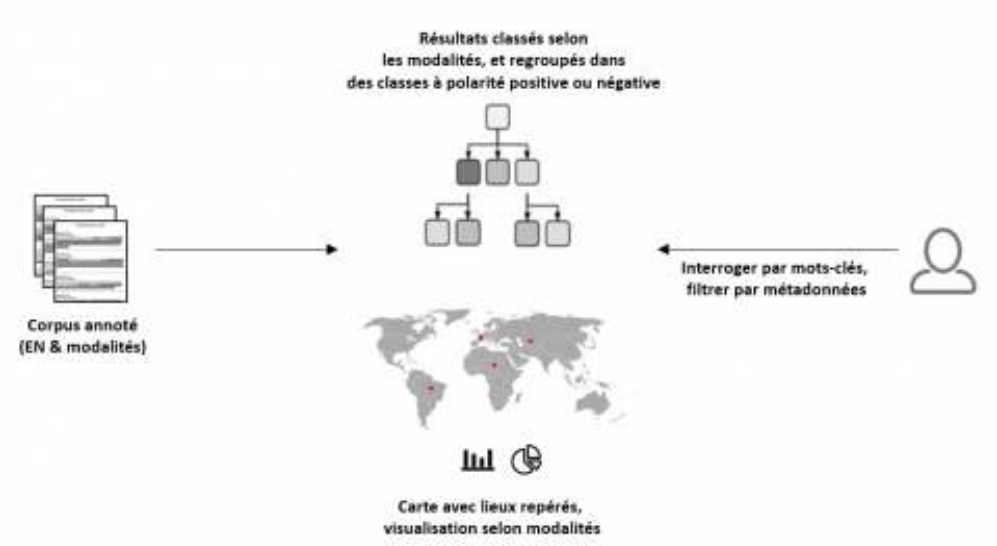

Image produite par les auteurs

82 D'autres fonctionnalités permettent de visualiser les résultats avec des statistiques sous forme graphique ou sous forme de concordancier, et de les exporter au format CSV.

83 Afin de faciliter la recherche dans l'interface, nous avons manuellement regroupé, à partir des gazetiers, toutes les variantes des EN spatiales autour d'une seule entrée. Cela permet de récupérer pour une requête donnée tous les résultats contenant les autres variantes historiques du mot recherché (par exemple باريس, پاريس, باريز pour باريس/Paris ; et فرانسا, المملكة الفرنساوية, بلاد الفرنسيس, فرنسا pour فرنسا /France, etc.).

Discussion et perspectives

84 Nous avons présenté une méthodologie pour une analyse pluridisciplinaire des entités nommées de lieu à Paris dans différents récits de voyage d'écrivains arabes. Notre contribution, particulièrement pour la langue arabe, réside dans l'analyse des modalités (opinions, sentiments et émotions) associées aux entités nommées de lieu, et distinguées par des polarités positive ou négative, ainsi que dans la représentation cartographique de ces lieux.

85 Les interprétations tirées de l'analyse des résultats montrent le grand intérêt de tels traitements automatiques. En effet, ces résultats confirment en très grande partie ce que les études critiques avancent sur l'impact positif de la découverte de Paris par les écrivains arabes, malgré une division conflictuelle qui apparaît dans leurs écrits.

86 Notre premier objectif est de systématiser progressivement ce travail en améliorant la REN et en complétant la représentation de Paris à partir de nouvelles sources annotées. Cela nous permettra ensuite d'envisager de nouvelles analyses sur les comparaisons, fréquentes dans les textes, entre la France et les pays d'origine des écrivains.

87 Nous envisageons également d'élargir la couverture des entités nommées vers le repérage des événements. En effet, on retrouve, dans le contexte des entités nommées de lieu, des mentions d'événements auxquels participent les personnages : rencontres (لقاء, موعد, تعارف), déplacements (وصل, سافر, مكث), participations à des actions, déclarations, etc. Ces passages méritent une attention particulière et contribueraient à l'analyse spatiale.

88 L'approche que nous avons adoptée pour l'annotation sémantique permet d'afficher la tonalité générale des passages (catégorisation sémantique fine et polarité). Même si cela nous offre déjà un outil puissant pour l'exploration des textes littéraires, nous estimons qu'une étape supplémentaire d'analyse des passages annotés serait nécessaire pour relier la modalité à son objet (sa cible). Bien souvent, il est nécessaire d'aller au delà d'une analyse syntaxique afin de déterminer la cible sémantique d'une modalité.

89 Enfin, en ce qui concerne la cartographie des lieux, nous étudions la faisabilité d'une implémentation de cartes abstraites et la spatialisation des lieux dans des espaces virtuels où les distances et les proportions ne représentent pas une cartographie à métrique euclidienne [32].

Bibliographie

Abdelali, Ahmed, Kareem Darwish, Nadir Durrani et Hamdy Mubarak. 2016. « Farasa : A Fast and Furious Segmenter for Arabic ». Dans *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, 11-16. San Diego : Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3003.

Alanazi, Saad. 2017. « A Named Entity Recognition System Applied to Arabic Text in the Medical Domain ». Thèse de doctorat, Staffordshire University.

Alotaibi, Fahd et Mark Lee. 2014. « A Hybrid Approach to Features Representation for Fine-grained Arabic Named Entity Recognition ». Dans *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, 984-995.

Dublin : Dublin City University and Association for Computational Linguistics. https://www.aclweb.org/anthology/C14-1093.

Alrahabi, Motasem. 2010. « Excom-2 : plateforme d'annotation automatique de catégories sémantiques : conception, modélisation et réalisation informatique : applications à la catégorisation des citations en arabe et en français ». Thèse de doctorat, université Paris 4. http://www.theses.fr/2010PA040005.

Alrahabi, Motasem. 2015. « E-Quotes : Enunciative Modalities Analysis Tool for Direct Reported Speech in Arabic ». Dans *Computational Linguistics and Intelligent Text Processing*, édité par Alexander Gelbukh, 479-490. Cham : Springer. https://doi.org/10.1007/978-3-319-18111-0_36.

Al-Sheikh, Khalil. 1998. *Paris in Modern Arabic Literature*. Amman, Jordan : Arab Establishment for Studies and Publishing.

Aly Mohamed Aly, Dalia. 2012. « Takhlis al-Ibriz de Rifa'a al-Tahtawy et sa traduction "L'Or de Paris" de Anouar Louca. Étude critique et approche linguistique ». Thèse de doctorat, université Paris III.

Aron, Paul, Laurence Brogniez, Tatiana Debroux, Jean-Michel Decroly et Christophe Loir. 2017. « À la recherche des chronotopes du roman urbain. Une cartographie des *Mystères de Bruxelles* (1845-1846) ». *Mappemonde* 121. https://doi.org/10.4000/mappemonde.3592.

Bakhtine, Mikhaïl. 1978. *Esthétique et théorie du roman*. Paris : Gallimard.

Bally, Charles. 1932. *Linguistique générale et linguistique française*. Berne : Francke.

Ben Mesmia, Fatma, Kais Haddar, Nathalie Friburger et Denis Maurel. 2017. « CasANER : Arabic Named Entity Recognition Tool ». Dans *Intelligent Natural Language Processing : Trends and Applications,* édité par Khaled Shaalan, Aboul Ella Hassanien et Fahmy Tolba, 173-98. Cham : Springer.

Benajiba, Yassine, Rosso Paolo, Miguel José et Ruiz Benedi. 2007. « ANERsys : An Arabic Named Entity Recognition System Based on Maximum Entropy ». Dans *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*. Berlin.

Boeglin, Noémie. 2018. « Représentations romanesques de la modernité parisienne dans le "Grand XIX^e^ siècle", 1830-1913 ». Thèse de doctorat, université de Lyon.

Borin, Lars, Dana Dannélls et Leif-Jöran Olsson. 2014. « Geographic Visualization of Place Names in Swedish Literary Texts ». *Literary and Linguistic Computing* 29 (3) : 400-404. https://doi.org/10.1093/llc/fqu021.

Boullier, Dominique et Audrey Lohard. 2012. *Opinion Mining et Sentiment Analysis : Méthodes et outils*. Marseille : OpenEdition Press. http://books.openedition.org/oep/198.

Cooper, David et Ian Gregory. 2011. « Mapping the English Lake District : A Literary GIS ». *Transactions of the Institute of British Geographers* 36 (1) : 89-108. https://doi.org/10.1111/j.1475-5661.2010.00405.x.

Culioli, Antoine. 1999. *Pour une linguistique de l'énonciation, formalisation et opérations de repérage*. Paris : Orphys.

Darwish, Kareem. 2013. « Named Entity Recognition Using Cross-Lingual Resources : Arabic as an Example ». Dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1558-1567. Sofia : Association for Computational Linguistics. https://www.aclweb.org/anthology/P13-1153.

Desclés, Jean-Pierre et Zlatka Guentcheva. 2000. « Énonciateur, locuteur, médiateur ». Dans *Les Rituels du dialogue*, édité par Aurore Monod Becquelin et Philippe Erikson, 79-112. Nanterre : Société d'ethnologie.

Dichy, Joseph. 2004. « Spécificateurs engendrés par les traits [±ANIME], [±HUMAIN], [±CONCRET] et structures d'arguments en arabe et en français ». Dans *De la mesure dans les termes. Hommage à Philippe Thoiron*, édité par Henri Béjoint et François Maniez, 151-181. Lyon : Presses universitaires de Lyon. https://halshs.archives-ouvertes.fr/halshs-00392582.

Does, Jesse de, Katrien Depuydt, Karen van Dalen-Oskam et Maarten Marx. 2017. « Namescape : Named Entity Recognition from a Literary Perspective ». Dans *CLARIN in the Low Countries*, édité par Jan Odijk et Arjan van Hessen, 361-370. London : Ubiquity Press. https://doi.org/10.5334/bbi.30.

Dominguès, Catherine, Laurence Jolivet, Carmen Brando et Marion Cargill. 2019. « Place and Sentiment-Based Life Story Analysis : From the Spanish Republican Army to the French Resistance ». *Revue française des sciences de l'information et de la communication* 17 (août). https://doi.org/10.4000/rfsic.7228.

Dupont, Yoann. 2017. « Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique ». Communication présentée à *TALN 2017*. Orléans, 26-30 juin. https://hal.archives-ouvertes.fr/hal-02448614.

Esuli, Andrea et Fabrizio Sebastiani. 2006. « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining ». Dans *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association.

Etaiwi, Wael, Arafat Awajan et Dima Suleiman. 2017. « Statistical Arabic Name Entity Recognition Approaches : A Survey ». *Procedia Computer Science* 113 (janvier) : 57-64. https://doi.org/10.1016/j.procs.2017.08.288.

Frontini, Francesca, Carmen Brando et Jean-Gabriel Ganascia. 2016. « REDEN ONLINE : Disambiguation, Linking and Visualisation of References in TEI Digital Editions ». Communication présentée à *Digital Humanities 2016.* Cracovie, 11-16 juillet. https://hal.archives-ouvertes.fr/hal-01395125.

Frontini, Francesca, Carmen Brando, Marine Riguet, Clémence Jacquot et Vincent Jolivet. 2016. « Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data ». *MALTIT : Materialities of Literature* 4 (2) (juillet). https://doi.org/10.14195/2182-8830_4-2_3.

Gabay, Simon et Giovanni Pietro Vitali. 2019. « A Theatre of Places : Mapping 17th French Theatre ». Dans *GIR'19 – Proceedings of the 13th Workshop on Geographic Information Retrieval*. ACM Digital Library. https://doi.org/10.1145/3371140.3371146.

Habash, Nizar. 2010. *Introduction to Arabic Natural Language Processing*. Williston : Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010.

Helwe, Chadi et Shady Elbassuoni. 2019. « Arabic Named Entity Recognition via Deep Co-Learning ». *Artificial Intelligence Review* 52 (1) : 197-215. https://doi.org/10.1007/s10462-019-09688-6.

Jonasson, Kerstin. 1994. *Le Nom propre : constructions et interprétations*. Louvain-la-Neuve : Éditions Duculot.

Khatib, Randa El, Julia El Zini, David Wrisley, Mohamad Jaber et Shady Elbassuoni. 2016. « TopoText : Interactive Digital Mapping of Literary Text ». Dans *Proceedings of COLING 2016 – 26th International Conference on Computational Linguistics : System Demonstrations,* 189-193.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho et Iryna Gurevych. 2018. « The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation ». Dans *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations,* 5-9. Santa Fe : Association for Computational Linguistics.

Lagrange, Frédéric. 2009. « Les femmes de Paris vues par trois voyageurs arabes du XIX^e^ siècle ». Communication présentée à *Connaissance de l'Orient*. Abu Dhabi, 13-14 janvier.

Leidner, Jochen Lothar. 2008. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Boca Raton : Universal Publishers.

Liu, Bing. 2010. « Sentiment Analysis and Subjectivity ». Dans *Handbook of Natural Language Processing*, édité par Nitin Indurkhya et Fred J. Damerau. CRC Press.

Mesfar, Slim. 2008. « Analyse morphosyntaxique automatique et reconnaissance des entités nommées en arabe standard ». Thèse de doctorat, université de Franche-Comté.

Moncla, Ludovic et Mauro Gaio. 2015. « A Multi-Layer Markup Language for Geospatial Semantic Annotations ». Dans *GIR'15 – Proceedings of the 9th Workshop on Geographic Information Retrieval*. ACM Digital Library. https://doi.org/10.1145/2837689.2837700.

Moncla, Ludovic, Mauro Gaio et Thierry Joliveau. 2018. « Cartographier les odonymes de Paris cités dans les romans du XIX^e^ siècle ». Communication présentée à *Atelier Humanités numériques spatialisées (HumaNS'2018) – SAGEO 2018*. Montpellier, 6 novembre.

Moretti, Franco. 1999. *Atlas of the European Novel, 1800-1900*. Londres : Verso.

Murrieta-Flores, Patricia, Christopher Donaldson et Ian Gregory. 2017. « GIS and Literary History : Advancing Digital Humanities research Through the Spatial Analysis of Historical Travel Writing and Topographical Literature ». *Digital Humanities Quarterly* 11 (01). http://www.digitalhumanities.org/dhq/vol/11/1/000283/000283.html.

Nadeau, David et Satoshi Sekine. 2007. « A Survey of Named Entity Recognition and Classification ». *Lingvisticæ Investigationes* 30 (janvier). https://doi.org/10.1075/li.30.1.03nad.

Nouvel, Damien, Maud Ehrmann et Sophie Rosset. 2015. *Les Entités nommées pour le traitement automatique des langues*. Londres, ISTE Éditions. https://hal-inalco.archives-ouvertes.fr/hal-01359438.

Oudah, Mai et Khaled Shaalan. 2012. « A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach ». Dans *24th International Conference on Computational Linguistics – Proceedings of COLING 2012 : Technical Papers*.

Pang, Bo et Lillian Lee. 2008. « Opinion Mining and Sentiment Analysis ». *Foundations and Trends in Information Retrieval* 2 (1-2) : 1-135. https://doi.org/10.1561/1500000011.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow et Ryan Roth. 2014. « Madamira : A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic ». Dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.

Piatti, Barbara, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni et William Cartwright. 2009. « Mapping Literature : Towards a Geography of Fiction ». Dans *Cartography and Art*, édité par William Cartwright, Georg Gartner et Antje Lehn, 1-16. Berlin, Heidelberg : Springer. https://doi.org/10.1007/978-3-540-68569-2_15.

Piolat, Annie et Rachid Bannour. 2009. « EMOTAIX : un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif ». *L'Année psychologique* 109 (4) : 655-698. https://doi.org/10.4074/S0003503309004047.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton et Christopher D. Manning. 2020. « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages ». Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, 101-108. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Schiller, Benjamin, Johannes Daxenberger et Iryna Gurevych. 2020. « Stance Detection Benchmark : How Robust Is Your Stance Detection ? ». *arXiv:2001.01565 [cs]*, janvier. http://arxiv.org/abs/2001.01565.

Shaalan, Khaled. 2014. « A Survey of Arabic Named Entity Recognition and Classification ». *Computational Linguistics* 40 (2) : 469-510. https://doi.org/10.1162/COLI_a_00178.

Shaalan, Khaled et Hafsa Raza. 2009. « NERA : Named Entity Recognition for Arabic ». *Journal of the American Society for Information Science and Technology* 60 (8) : 1652-1663. https://doi.org/10.1002/asi.21090.

Sharnagat, Rahul. 2014. « Named Entity Recognition : A Literature Survey ». Center For Indian Language Technology.

Soudani, Aicha, Yosra Meherzi, Asma Bouhafs, Francesca Frontini, Carmen Brando, Yoann Dupont et Frédérique Mélanie-Becquet. 2019. « Adapting a System for Named Entity Recognition and Linking for 19th Century French Novels ». Communication présentée à *Digital Humanities Conference, ADHO*. Utrecht, 9 juillet. https://dh-abstracts.library.cmu.edu/works/9768.

Turney, Peter D. 2002. « Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews ». Dans *ACL'02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 417-424. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073153.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell et Melanie Martin. 2004. « Learning Subjective Language ». *Computational Linguistics* 30 (3) : 277-308. https://doi.org/10.1162/0891201041850885.

Zhang, Lei. 2012. « Analyse automatique d'opinion : problématique de l'intensité et de la négation pour l'application à un corpus journalistique ». Thèse de doctorat, université de Caen. https://tel.archives-ouvertes.fr/tel-00777603.

Zhang, Lei et Bing Liu. 2017. « Sentiment Analysis and Opinion Mining ». Dans *Encyclopedia of Machine Learning and Data Mining*, édité par Claude Sammut et Geoffrey I. Webb, 1152-1161. Boston : Springer. https://doi.org/10.1007/978-1-4899-7687-1_907.

Notes

1 Plus récemment, le projet *Open Islamicate Texts Initiative* (https://github.com/OpenITI) vise à encoder et éditer une masse importante de textes islamiques prémodernes, l'objectif étant de pouvoir exploiter les corpus avec des outils d'analyse textuelle. Le genre et le domaine cible des textes sont variés (religieux, médical…) et, pour l'instant, la question du genre littéraire a reçu peu d'attention.

2 https://www.lancaster.ac.uk/chronotopic-cartographies/.

3 https://www.lancaster.ac.uk/chronotopic-cartographies/visualisations/frankenstein/deep-chronotopes/.

4 Aron *et al.* (2017) soulignent que *Les Mystères de Bruxelles* relève des multiples imitations des *Mystères de Paris* d'Eugène Sue.

5 http://www.literaturatlas.eu.

6 Attribution de coordonnées à un nom de lieu afin de pouvoir le projeter sur une carte géographique.

7 Le référentiel géohistorique mobilisé est issu de GeoHistoricalData : http://geohistoricaldata.org.

8 https://github.com/cvbrandoe/ArabicLitTAL/blob/master/annotated-data-original/Guide%20Ester.pdf.

9 https://wiki.dbpedia.org.

10 https://www.geonames.org.

11 https://www.historypin.org/en/victorian-london/.

12 http://www.lancaster.ac.uk/mappingthelakes/.

13 Le corpus provient du dépôt gratuit de la fondation Hindawi : http://www.hindawi.org.

14 La revue satirique en langue arabe parlée égyptienne de Jacob Sinoua *Abou Nazzâra, ou L'Homme aux lunettes* (1839-1912) était imprimée à Paris et circulait sous le manteau en Égypte. C'est que Paris n'est pas seulement un lieu de science ou de formation, c'est aussi un lieu de pensée révolutionnaire. En 1884, c'est aussi à Paris que les deux grands réformistes musulmans Jamal Eddine Al-Afghânî et Muhammed Abdû fondent la société secrète Al-ʿUrwa l-wuthqâ (« L'anse la plus solide ») et publient la revue *L'Anse la plus solide et la grande révolution historique*, qui sera interdite de diffusion en Égypte et en Inde par les autorités britanniques.

15 Un répertoire de noms de lieux avec une localisation directe associée, souvent sous forme de coordonnées ; les données sont souvent stockées dans une base de données et mises à disposition en ligne.

16 http://www.ign.fr/institut/activites/referentiel-a-grande-echelle/.

17 https://pleiades.stoa.org.

18 Pour un historique des campagnes d'évaluation dédiées aux EN, se référer à Nouvel, Ehrmann et Rosset 2015 (p. 65).

19 https://inception-project.github.io.

20 https://github.com/cvbrandoe/ArabicLitTAL/blob/master/annotated-data-original/Guide%20Ester.pdf.

21 Pour les besoins de l'analyse, nous avons choisi d'annoter dès le départ toute EN spatiale, qu'elle soit en France ou ailleurs. Les auteurs comparent souvent dans leur récit la France et leur pays d'origine. Cette couverture sera donc très importante pour le projet à long terme. Cependant, nous ne prenons en compte dans cet article que les EN appartenant à la France.

22 https://github.com/cvbrandoe/ArabicLitTAL/tree/master/annotated-data-original/.

23 La procédure d'évaluation de performances d'un système NER correspond à celle proposée par la tâche 9.1 de la campagne d'évaluation SemEval 2013 (code source : https://github.com/davidsbatista/NER-Evaluation/).

24 https://fr.wikipedia.org/wiki/Précision_et_rappel.

25 À ce titre, une trentaine de récits de voyage dont les événements se passent à Paris et en France sont en cours de numérisation dans le cadre du projet.

26 Voici un exemple de règle élémentaire : « Si dans une phrase on trouve une EN de lieu, et si dans le contexte de cette EN on trouve un marqueur de modalité qui ne soit pas entouré de négation, alors annoter la phrase avec l'étiquette en question. »

27 Notons que les cibles ou objets des modalités (Liu 2010) dans les 36 bonnes annotations renvoient à des EN en France, puisque les phrases évaluées contiennent nécessairement ce type d'EN.

28 https://frama.link/ArabicMapping.

29 L'analyse critique d'Al-Sheikh est confirmée par deux critiques arabes très influents : Ihsan Abbas, qui a écrit l'introduction du livre d'Al-Sheikh, et Jaber Asfour, qui a révisé le même livre pour le magazine Al-Arabi.

30 Cette ambivalence est très bien résumée dans les vers d'al-Tahtawi (traduits par J. Dichy) :
أيوجد مثل باريسَ ديار / شموس العلم فيها لا تغيب
وليل الكفر ليس له صباح / أما هذا وحقّكم عجيب؟
Existe-t-il séjour comme Paris / où le soleil des sciences règne sans une absence
Où nul matin pourtant ne dissipe la nuit / du refus de la vraie croyance ?
Rien au monde ne laisse, croyez-moi, plus surpris !

31 https://obvil.huma-num.fr/ariane/humanistica/search/.

32 https://www.lancaster.ac.uk/chronotopic-cartographies/visualisations/frankenstein/topoi/.

Auteurs

Motasem Alrahabi

Sorbonne Université, Paris, France
Motasem Alrahabi est ingénieur de recherche à Sorbonne Université. Ses recherches dans le domaine des humanités numériques et du traitement automatique des langues portent principalement sur l'analyse sémantique et discursive de textes en français et en arabe.
ORCID 0000-0001-5478-4283
motasem.alrahabi@paris-sorbonne.fr

Carmen Brando

École des hautes études en sciences sociales, Paris, France
Carmen Brando (PhD) est ingénieure de recherche en humanités numériques à l'EHESS, spécialiste des méthodes en traitement automatique des langues, du Web sémantique et de la géomatique appliquée aux sciences humaines et sociales.
ORCID 0000-0001-7098-3522
carmen.brando@ehess.fr

Muhamed AlKhalil

Université de New York, Abou Dabi, Émirats arabes unis
Muhamed AlKhalil est professeur d'arabe et fondateur du programme d'études arabes à NYUAD. Ses intérêts de recherche portent sur les interactions entre le littéraire et le politique dans la littérature arabe moderne. Il est également responsable du projet SAMER (Simplification of Arabic Masterpieces for Extensive Reading).
muhamed.alkhalil@nyu.edu

Joseph Dichy

Université canadienne de Dubaï, Dubaï, Émirats arabes unis
Joseph Dichy est professeur à l'université canadienne de Dubaï (CUD) et directeur de collection à AraDic-Monde arabe Éditions (Lyon). Ses recherches portent sur la linguistique et le traitement automatique de l'arabe, la lexicologie bilingue et la rhétorique arabe médiévale.
ORCID 0000-0002-9123-7358
joseph.dichy@cud.ac.ae

Droits d’auteur

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.